A Generic Multivariate Distribution for Counting Data

Marcos Capistrán* and J. Andrés Christen†
Centro de Investigación en Matemáticas, A. C. (CIMAT)
Guanajuato, MEXICO.

March 2011

Abstract

Motivated by the need, in some Bayesian likelihood free inference problems, of imputing a multivariate counting distribution based on its vector of means and variance-covariance matrix, we define a generic multivariate discrete distribution. Based on blending the Binomial, Poisson and Negative-Binomial distributions, and using a normal multivariate copula, the required distribution is defined. This distribution tends to the Multivariate Normal for large counts and has an approximate pmf version that is quite simple to evaluate.

KEYWORDS: Counting Data; Bayesian inference in epidemics; Copulas.

1 Introduction

We develop a generic discrete multivariate distribution defined in terms of its mean and covariate matrix only. The multivariate Normal distribution is defined in such terms and would be the default option as an inputed distributions when only the mean and covariance matrix are available for an

*Email: *marcos@cimat.mx* .
†(corresponding author) Centro de Investigación en Matemáticas, A. C. (CIMAT), A.P. 402, Guanajuato, Gto. 36000, Mexico. Tel: +52-(473)-732-7155. Fax: +52-(473)-732-5749. Email: *jac@cimat.mx* .

otherwise unknown distribution. However, there is no alternative when considering discrete data, specially in the case of low counts where a Normal approximation is not feasible. The motivation to defining such discrete distribution is as follows.

The development and analysis of mathematical epidemic models that take into account uncertainty is an active field of research (Bretó et al., 2009; Alonso et al., 2007; Finkenstadt et al., 2002; Schwartz et al., 2004; Nåsell, 2002; Chen and Bokka, 2005; Andersson and Britton, 2000). The importance of this field of research is apparent given its potential impact on public health policies to handle emergent and re-emergent infectious diseases such as dengue fever, Lyme disease, tuberculosis, flu, etc. It is known that the effects of local (demographic) stochasticity weight more in determining the dynamics of epidemics when the number of individuals in the population is low. On the other hand, parameter estimation is among the standard tools to explore the predictive capacity of models from partial observation of the state variables. In this context, it is specially important to devise methods to study the predictive capacity of the mathematical models, in particular, to quantify the uncertainties.

For the sake of clarity we use the simplest epidemic model, e.g. the SIR model without vital dynamics. Let the random variables X_1, X_2 and X_3 denote the number of Susceptible, Infected and Recovered individuals in a closed population at a given time t, respectively. The stochastic model is defined by the processes through which it evolves:

$$X_1 + X_2 \xrightarrow{b_0 \frac{X_2}{\Omega}} 2X_2$$
$$X_2 \xrightarrow{b_1} X_3$$

If we denote by x_1, x_2 and x_3 a realization of the random variables X_1, X_2 and X_3, and let $P_{\bar{x}}(t) = P(x_1, x_2, x_3, t)$ be the probability that the system is in state (x_1, x_2, x_3) at time t, then the chemical master equation (van Kampen, 1992) for this system is given by

$$\begin{aligned} \frac{dP(x_1, x_2, x_3, t)}{dt} =& b_0 \frac{(x_1 + 1)}{\Omega}(x_2 - 1)P(x_1 + 1, x_2 - 1, x_3, t) \\ &+ b_1(x_2 + 1)P(x_1 + 1, x_2 - 1, x_3 - 1, t) \\ &- (b_0 \frac{x_1}{\Omega} x_2 + b_1 x_2)P(x_1, x_2, x_3, t), \end{aligned} \tag{1}$$

where the constant Ω denotes the total number of individuals, and b_0 and b_1 are respectively the contact rate and the rate of loss of infectiousness.

Applying the *Inverse Size Expansion* (van Kampen, 1992) to equation 1 leads to equations for the expected value and the fluctuations of x_1, x_2 and x_3. The Fokker-Plank approximation is then used leading to an approximation for the mean $E_t(X_i)$ and cross products $E_t(X_i X_j)$ for all species $i, j = 1, 2, \ldots, k$ at each time t Chen and Bokka (2005). As far as the distribution for the X_j's is concern we know to be dealing with counting data $X_i \in \mathbb{N}; i = 1, 2, \ldots, k$ and, for a fixed time t, we have (an approximation for) their means $\mu_i = E_t(X_i)$, variances $v_i = E_t(X_i^2) - \mu_i^2$ and their correlations $\rho_{ij} = \frac{E_t(X_i X_j) - \mu_i \mu_j}{\sqrt{v_i v_j}}$.

It is possible to simulate directly from the true model $P(\cdot, t)$ above (for fixed b_0 and b_1) to simulate realizations of $(X_1(t), X_2(t), X_3(t))$ (Gillespie, 2007). Data is commonly only available for $X_2(t)$ and for specific epidemics (eg. Dengue fiver) there is substantial prior expert information for the contact rate, b_0, and the rate of loss of infectiousness, b_1. It would be possible to use the ABC algorithm (Marin et al., 2011) to make Bayesian inferences about b_0 and b_1 but the simulation procedure becomes very slow for moderate population sizes Ω and still the ABC approach lacks a formal theoretical foundation (Marin et al., 2011, p. 4). Instead, using the moment approximation explained above (that indeed depends on the parameters of interest b_0 and b_1), we impute a counting (discrete) distribution on the observables (commonly $X_2(t)$ but in some situations all $X_1(t), X_2(t), X_3(t)$ are observed unknonw, 1978) matching those moments to create a likelihood. The computational complexity of this likelihood is in fact independent of Ω and, using elicited priors and MCMC, a Bayesian inference is possible for any set of population sizes. We have had already promising results along these lines and will publish such research in an specialized journal of the field.

We will assume that the correlation matrix $\boldsymbol{\rho} = (\rho_{ij})$ is positively defined and, certainly, $\mu_i, v_i > 0$. Based on this information only, we need to impute a discrete distribution for the observables (X_1, X_2, X_3) that would be defined by these moments. Here we propose such generic multivariate distribution for counting data. In the next section we explain the univariate version, which is a simple combination of the default distributions commonly used for counting data, namely, the Binomial, the Poisson and the Negative Binomial. In Section 3 we create the multivariate version using a Normal copula and in Section 4 present some examples.

2 A Univariate Generic Discrete Distribution $Gd(m, v)$

The Poisson, Binomial and Negative Binomial distributions are simple form distributions and first candidates for counting data models. For any mean and variance $\mu, v > 0$ we make a combination of these three distributions in the following way

$$Gd(x|\mu, v) = \begin{cases} C_x^m \; p^x(1-p)^{m-x}; \;\; x = 0, 1, \dots m & \text{if} \;\; \mu > v \\ e^{-v}\frac{v^x}{x!}; \;\; x \in \mathbb{N} & \text{if} \;\; \mu = v \\ C_{x-1}^{x+m-1} \; p^x(1-p)^m; \;\; x \in \mathbb{N} & \text{if} \;\; \mu < v \end{cases} \tag{2}$$

where $p = 1 - \min\left\{\frac{m}{v}, \frac{v}{m}\right\}$, $m = \frac{\mu^2}{|\mu - v|}$ and C_x^m are the combinations of m items taken in subsets of size x. That is, we use a Binomial if $\mu > v$, a Poisson if $\mu = v$ and a Negative-Binomial if $\mu < v$. Neither of these distributions can handle *any* mean and variance; by combining these distributions we obtain the Generic Discrete class $Gd(\mu, v)$ defined for arbitrary mean μ and variance v, $\mu, v > 0$, and these two moments completely define the distribution.

Indeed, it is straightforward to see that if $X \sim Gd(\mu, v)$, $E(X) = \mu$ and $V(X) = v$. More importantly, for a fixed mean μ, given both the properties of the Binomial and the Negative-Binomial, we see that

$$\lim_{v \to \mu} Gd(x|\mu, v) = e^{-v}\frac{v^x}{x!}.$$

Therefore we have a continuous evolution of this parametric class, being the Poisson the "continuous bridge" between the Binomial ($\mu > v$) and Negative Binomial ($\mu < v$). (Note that if $\mu > v$ and $v \to \mu$, the support will increase to cover all $\mathbb{N}$ since $m \to \infty$.)

Moreover, if $X \sim Gd(\mu, v)$, $\frac{X-\mu}{\sqrt{v}}$ will tend to a standard Normal distribution if $\mu \to \infty$ and $p \to p_0 \in (0, 1)$. That is, for large μ (and for example $\mu - 3\sqrt{v} > 0$) $Gd(\mu, v)$ can be approximated with a $N(\mu, v)$. Practical guidelines for approximating the Poisson, Binomial and Negative-Binomial distributions should be used when calculating the cdf, pmf etc. of $Gd(\mu, v)$. We then see that the $Gd(\mu, v)$ family evolves to a Normal distribution when μ is large (large counts).

3 The Multivariate Case

Suppose we have $k = 2$ discrete distributions and we let $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and $\boldsymbol{v} = (v_1, v_2)'$ be their vector of means and variances and $\rho = \rho_{1,2}$. We require a bivariate discrete distribution, defined in terms of $\boldsymbol{\mu}$, $\boldsymbol{v}$ and ρ, such that the marginal distribution for each X_i is $Gd(\mu_i, v_i)$ and the resulting correlation between X_1 and X_2 is (at least approximately) ρ.

We use a Normal Copula (see Nelsen, 2006, chap. 2) to create a a joint distribution. Let

$$\phi_\rho(s,t) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ -\frac{(s^2 - 2\rho st + t^2)}{2(1-\rho^2)} \right\}$$

be the bivariate standard normal distribution with correlation ρ and $\Phi(s)$ the standard normal cdf. The normal Copula is defined as

$$C_\rho(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \phi_\rho(s,t) ds dt.$$

We define the joint cdf of X_1 and X_2 as

$$F_{\boldsymbol{\mu},\boldsymbol{v},\rho}(x_1, x_2) = C_\rho(F_{\mu_1,v_1}(x_1), F_{\mu_2,v_2}(x_2)),$$

where $F_{\mu,v}$ is the cdf of a $Gd(\mu, v)$. $F_{\boldsymbol{\mu},\boldsymbol{v},\rho}(x_1, x_2)$ defines the Generic Discrete distribution in dimension 2, $Gd_2(\boldsymbol{\mu}, \boldsymbol{v}, \rho)$, and it is straightforward to verify that its pmf is

$$\begin{aligned} Gd_2(x_1, x_2|\boldsymbol{\mu}, \boldsymbol{v}, \rho) \;=\; & C_\rho(F_{\mu_1,v_1}(x_1), F_{\mu_2,v_2}(x_2)) - C_\rho(F_{\mu_1,v_1}(x_1 - 1), F_{\mu_2,v_2}(x_2)) - \\ & C_\rho(F_{\mu_1,v_1}(x_1), F_{\mu_2,v_2}(x_2 - 1)) + C_\rho(F_{\mu_1,v_1}(x_1 - 1), F_{\mu_2,v_2}(x_2 - 1)) \end{aligned} \tag{3}$$

(that is, the pmf are the corresponding jumps in the stepped cdf $F_{\boldsymbol{\mu},\boldsymbol{v},\rho}(x_1, x_2)$). Since $C_\rho(u, v)$ is a Copula for every $-1 \leq \rho \leq 1$, the marginal distributions will be precisely $Gd(\mu_1, v_1)$ and $Gd(\mu_2, v_2)$, as required (Nelsen, 2006).

By pretending $F_{\mu_i,v_i}(x_i)$ is differentiable, with derivative $Gd(x_i|\mu_i, v_i)$, 'differentiating' $F_{\boldsymbol{\mu},\boldsymbol{v},\rho}(x_1, x_2)$ suggest the following approximation to the pmf

$$Gd_2(x_1, x_2|\boldsymbol{\mu}, \boldsymbol{v}, \rho) \approx gd_2(x_1, x_2|\boldsymbol{\mu}, \boldsymbol{v}, \rho) = K \frac{\phi_\rho(s,t)}{\phi(s)\phi(t)} Gd(x_1|\mu_1, v_1) Gd(x_2|\mu_2, v_2) \tag{4}$$

were $s = \Phi^{-1}(F_{\mu_1,v_1}(x_1))$, $t = \Phi^{-1}(F_{\mu_2,v_2}(x_2))$ and $\phi(\cdot)$ is the standard Normal pdf, for some normalization constant K. This approximation will prove useful even for μ_i small, and is far less computationally demanding than the exact version in (3).

If $F_{\mu_i,v_i}(x_i)$ were cdf's of $N(\mu_i, v_i)$, that is $F_{\mu_i,v_i}(x_i) = \Phi\left(\frac{x_i-\mu_i}{\sqrt{v_i}}\right)$, we immediately see that the correlation between X_1 and X_2 is ρ. In the general case as each marginal distribution $Gd(\mu_i, v_i)$ becomes similar to a Normal distribution the actual correlation will be approximately ρ and Gd_2 becomes increasingly similar to a Bivariate Normal distribution with the correct moments.

Finally, the multivariate generic discrete distribution is defined as

$$Gd_n(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{v}, \boldsymbol{\rho}) = \int_{\Phi^{-1}(F_{\mu_1,v_1}(x_1-1))}^{\Phi^{-1}(F_{\mu_1,v_1}(x_1))} \cdots \int_{\Phi^{-1}(F_{\mu_n,v_n}(x_n-1))}^{\Phi^{-1}(F_{\mu_n,v_n}(x_n))} (2\pi)^{-n/2}|\boldsymbol{\rho}|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{s}'\boldsymbol{\rho}^{-1}\boldsymbol{s}\right\} ds_1 \cdots ds_n.$$

The corresponding approximate pmf would be

$$gd_n(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{v}, \boldsymbol{\rho}) = K\frac{\phi_{\boldsymbol{\rho}}(s_1, s_2, \ldots, s_n)}{\phi(s_1)\phi(s_2)\cdots\phi(s_n)} Gd(x_1|\mu_1, v_1)Gd(x_2|\mu_2, v_2)\cdots Gd(x_n|\mu_n, v_n),$$

with $s_i = \Phi^{-1}(F_{\mu_i,v_i}(x_i))$.

4 Example

We present four examples of Gd_2, see Figure 1 and Table 1. We compare the approximation with the exact distributions by calculating numerically the exact pmf in (3) and the approximate pmf, gd_2, in (4) over a relevant grid of the support for each example. The approximation seems to be very good option and far less computationally demanding. Moreover, the moments match correctly and both Gd_2 and gd_2 have an actual correlation quite near the required one (compare ρ with ρ' and ρ^* in Figure 1). It is also very remarkable that the normalization constant needed for the approximation is quite close to 1. This will potentially enable the use of gd_n as an alternative, less computationally demanding likelihood, by considering K to depend only marginally on the mean and variance-covariance matrix.

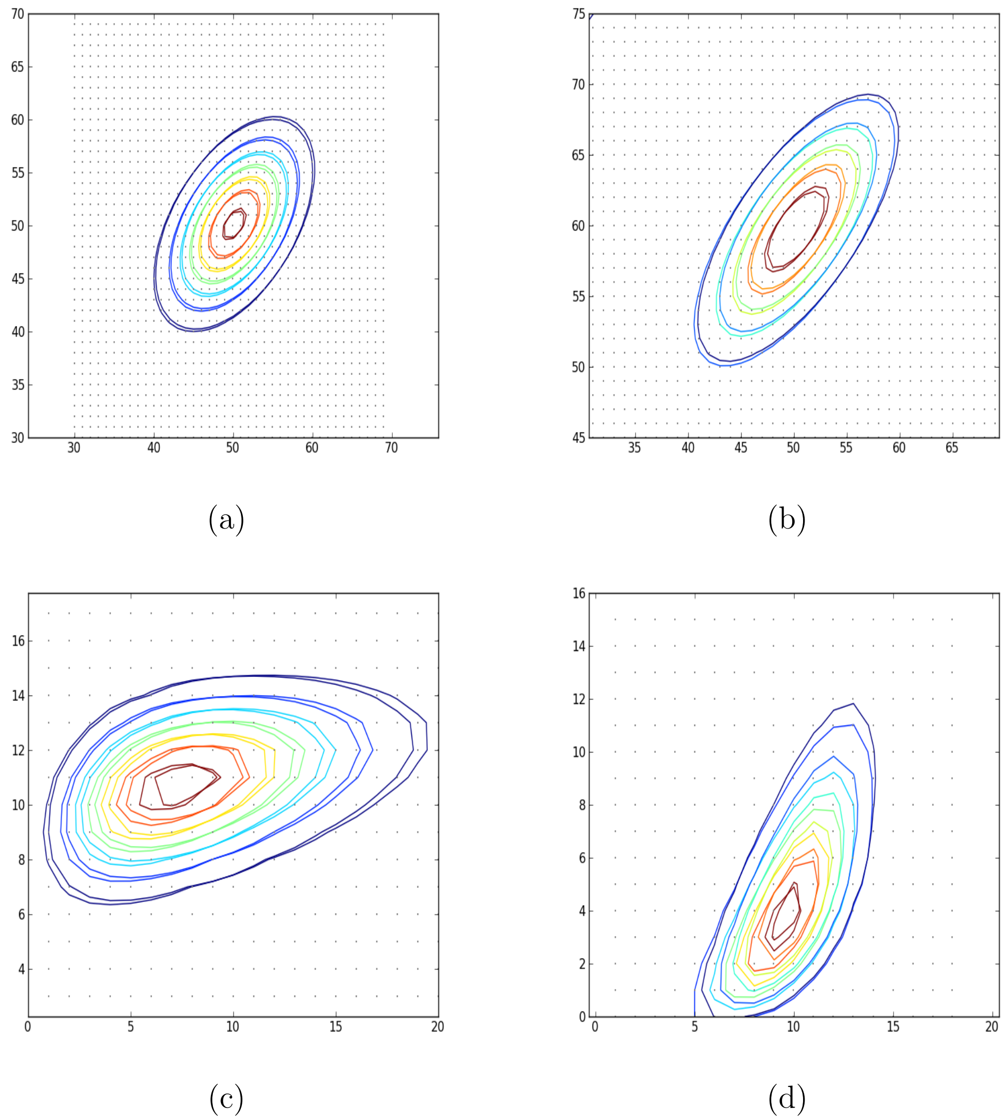

Figure 1: Contour plots for the exact and approximate pmf's of Gd_2, for the parameters described in Table 1. The contours in all cases basically match.

	μ_1	v_1	μ_2	v_2	ρ	ρ'	K	μ_1^*	v_1^*	μ_2^*	v_2^*	ρ^*
(a)	50	25	50	25	0.5	0.5144	0.99	50.25	24.99	50.25	24.99	0.5009
(b)	50	25	60	25	0.7	0.7129	0.99	50.35	25.04	59.85	24.76	0.7013
(c)	10	25	11	4	0.4	0.3988	0.99	9.45	23.30	10.90	3.85	0.3876
(d)	10	5	5	10	0.7	0.6869	0.97	10.23	4.68	5.43	10.43	0.6670

Table 1: Gd_2, resulting correlation, ρ', for the exact and resulting moments for the approximate (*) pmf of Gd_2, gd_2, with various parameters. Contour plots for the exact and approximate pmf's of distribution (a)-(d) may be seen in Figure 1. The normalization constant needed for gd_2 is given in column K.

5 Discussion

We develop a generic discrete multivariate distribution defined in terms of its vector of means and variance-covariance matrix only, as it is the case for the Multivariate-Normal distribution for continuous data. This distribution has applications in the Bayesian analysis of complex models were we are dealing with counting data and the correct likelihood is not available analytically, but approximation techniques can be developed to obtain moments of observables. This is the case when studying epidemics using the SIR stochastic model, as explained in Section 1. The distribution developed here can now be used as a default distribution to be imputed to multivariate counting data in such situations. Moreover, when large counts are involved this distribution tends to a Multivariate Normal, (eg. Figure 1(a) and (b)).

References

Alonso, D., A. McKane, and M. Pascual (2007). Stochastic amplification in epidemics. *Journal of the Royal Society Interface 4*(14), 575.

Andersson, H. and T. Britton (2000). *Stochastic epidemic models and their statistical analysis*. Springer Verlag.

Bretó, C., D. He, E. Ionides, A. King, S. Ghosal, J. Lember, A. van der Vaart, K. Triantafyllopoulos, P. Harrison, M. Ribatet, et al. (2009). Time series analysis via mechanistic models. *Annals 3*(1), 319–348.

Chen, W. and S. Bokka (2005). Stochastic modeling of nonlinear epidemiology. *Journal of theoretical biology 234*(4), 455–470.

Finkenstadt, B., O. Bjornstad, and B. Grenfell (2002). A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. *Biostatistics 3*(4), 493.

Gillespie, D. (2007). Stochastic simulation of chemical kinetics. *Annual review of physical chemistry 58*(1), 35–55.

Marin, J., P. Pudlo, C. P. Robert, and R. Ryder (2011). Approximate Bayesian Computational methods. Technical report, http://adsabs.harvard.edu/abs/2011arXiv1101.0955M.

Nåsell, I. (2002). Stochastic models of some endemic infections. *Mathematical Biosciences 179*(1), 1–19.

Nelsen, R. (2006). *An introduction to copulas.* New York: Springer.

Schwartz, I., L. Billings, and E. Bollt (2004). Dynamical epidemic suppression using stochastic prediction and control. *Physical Review E 70*(4), 46220.

unknonw (1978). Influenza in a boarding school. *British Medical Journal 1*(6112), 587.

van Kampen, N. (1992). *Stochastic processes in physics and chemistry.* North Holland.